面向先秦典籍的知识本体构建技术研究*

■ 何琳 陈雅玲 孙珂迪
南京农业大学信息管理系 南京 210095

摘 要:[目的/意义]构建面向典籍文本的语义本体,能够促进典籍文本的挖掘与分析。然而由于典籍文本与现代文本在语法上存在较大差异,给面向典籍的语义本体构建带来了困难。[方法/过程]本文运用自然语言处理技术探讨针对先秦典籍的本体构建方法。以国际上文化遗产领域通用的 CIDOC CRM 为框架,设计先秦典籍本体模型。针对典籍文本内容的特点及句法特征,将规则抽取与条件随机场方法相结合,提出一套本体实例自动获取技术,并以《左传》为实验语料进行测试。[结果/结论]实验表明,本文所提出的本体实例抽取技术能够较好地提高面向典籍文本的本体构建效率。基于规则的本体实例抽取实验 F 值在93%左右,基于条件随机场的本体实例抽取最佳特征模板的 F 值为82.51%。在本体实例获取中,词性信息和位置信息具有重要作用。

关键词:先秦典籍 左传 本体构建 条件随机场 规则匹配
分类号:G254
DOI:10.13266/j.issn.0252-3116.2020.07.002

1 引言

中华文明五千年来一脉相承、从未中断,其中一个重要的原因就是中华民族拥有各个历史时期浩如烟海的典籍[1]。将蕴含在典籍中的知识形式化、模型化描述及关联关系揭示,不仅有利于与传统文化有关的知识的识别、理解和共享,而且有助于推动典籍文本的深度挖掘及人文计算研究的开展。20世纪前半叶,哈佛燕京学社引得编撰处编纂了书本式逐字词古籍索引《汉学引得丛刊》,将典籍知识的描述从文献单元深入至词汇单元。本体等语义网技术的发展为典籍文本的知识组织提供了新的活力。学界开展了诸如家谱资源描述本体模型[2]、"二十四史"本体[3]、中医古籍本体[4]等相关的研究。

总体来说,构建面向典籍的知识本体,虽然已有相关的研究,但仍然面临诸多的困难。一方面,仍然缺少面向典籍文本的顶层语义描述框架;另一方面,典籍的语法和句法同现代语言存在较大差异,针对典籍文本的概念及概念关系的挖掘方法仍然需要进一步深入研究。在此背景下,本文尝试通过复用已有本体设计具有较好通用性的本体语义框架,继而重点研究本体实例的获取方法。结合典籍文本的内容特点,笔者提出将规则抽取与条件随机场识别相结合,最大限度利用古汉语中的固有表述规则进行本体实例的抽取,并以《左传》为实验语料对本文提出的方法进行测试。

2 研究综述

近年来,典籍文本的数字化已经取得了较为丰硕的成果,在此基础上,学者们围绕古汉语文本的自动分词、词性标注、命名实体识别与词义研究等开展了许多探索性的研究,并取得了一批研究成果[5]。

分词方面,基于条件随机场模型的古文自动分词方法取得了较好的效果,以《孟子》《左传》《汉书》《诗经》等典籍为语料进行的相关实验表明,分词的最好结果 F 值可接近98%[6]。词性标注方面,学者们先后在《楚辞》《明史》《左传》《论语》等典籍文本中进行了相关的实验研究并取得了较好的效果,其中效果最好的实验调和平均数 F 值接近95%[7]。命名实体识别方面的研究成果比分词与词性标注方面的成果相对较少一些,有学者研究了从《孟子》《左传》《二十四史》《三国志》等典籍中抽取人名、地名及时间等命名实体的方法,还有学者研究了方志中地名、物产等名词的抽

* 本文系中央高校基本科研业务费资助项目"基于《汉学引得丛刊》的古文本体研究"(项目编号:SKCX2017004)研究成果之一。
作者简介:何琳(ORCID:0000-0002-4207-3588),教授,博士,博士生导师,E-mail:helin@njau.edu.cn;陈雅玲(OCRID:0000-0002-7515-4843),硕士研究生;孙珂迪(ORCID:0000-0003-0193-1117),硕士研究生。
收稿日期:2019-07-10 修回日期:2019-11-24 本文起止页码:13-19 本文责任编辑:王传清

取[4,8]。词义研究继承和利用了自动分词、词性标注的成果[9]，也为句法分析、语义标注提供了研究基础，学者们借助多种技术方法，实现了古汉语的词义消歧及少量浅层句法标注研究。

从上述研究可以看出，目前针对古文的自动分词和词性标注等取得了一定的成果，但如何更好地应用这些成果，在从典籍文本中抽取命名实体及实体关系的基础上，构建领域本体等语义化描述工具，推动典籍文本的深入挖掘，仍然需要开展更进一步的研究。

目前本体语义框架的建立都需要人工参与。如何在已建立的本体语义框架下，采用自然语言处理、机器学习等技术从语料中自动抽取相关术语及属性关系，完成本体实例的自动学习和获取是实现本体应用的重要途径[10]。近年来，本体自动构建的研究产生了一系列的成果。一些学者基于维基百科、WordNet 等资源构建了 DBPedia Ontology、YAGO 等大型通用本体。以生命科学、地理科学为代表的自然科学领域已经有较为大型的实用化领域本体，如 GeoNames Ontology、The Drug Ontology、UMLS SemNet、Gene Ontology 及 SNOMED 等[11]。而在人文社会科学领域中，部分学者在历史学、哲学等领域尝试开展诸如三国志本体、国共合作本体、国史本体、二十四史本体及哲学本体的构建研究[12-13]。上海图书馆针对馆藏家谱资源设计了本体模型[2]，中华书局主持开发了“二十四史”本体对人物、地点、时间等进行语义化组织[4]，中国中医科学院的中医古籍知识库则从中医古籍中抽取知识元并建立了知识元间的属性关系[3]。学者们还尝试将元数据、本体技术应用于戏剧、民俗等领域的信息资源描述和组织[14]。

总体来说，自然科学领域概念及概念之间的关系较为明确，大型的、实用化的领域本体自动构建发展较为迅速，而人文社会科学领域存在顶层语义框架难以界定、概念关系较为灵活等特性，大规模的、实用化的领域本体比较少见。因此，本文将针对典籍文本的内容，探讨建立规范的、移植性好的本体框架以及相应的本体实例自动获取方法与技术。

3 面向先秦典籍的本体模型构建

3.1 先秦典籍本体构建的难点

同已有的本体构建方法与技术相比，先秦典籍本体的构建存在以下两方面的难题。

3.1.1 本体语义框架的设计

先秦是中国传统思想和文化发源的重要时期，先秦典籍记录了中华民族思想理念、传统美德和人文精神等。因此，典籍中蕴含的知识点类别广泛且语义关系复杂。如何设计一个本体语义框架，对其中蕴含的知识进行形式化、模型化描述及关联关系揭示，这是先秦典籍本体构建所面临的难题之一。

在本体模型的设计方面，为了能够实现对先秦典籍中蕴含知识点的理解和共享，笔者参考了大量已有本体项目。其中，由国际工作委员会提出的面向对象的 CIDOC CRM 概念参考模型，用本体方式描述了文化遗产工作中所需要的概念体系及关系的定义和形式结构，以达到对文化遗产的共同理解。该模型已经被广泛应用于物质与非物质文化遗产领域[15-16]，具有较好的通用性。依据 CIDOC CRM 概念模型，笔者利用先秦典籍研究文献，对典籍的内容进行梳理[17-18]，并对典籍的内容进行聚类分析（参见本专题的《春秋时期社会发展的主题挖掘与演变分析》一文），确定了军事、婚姻、外交、政治、民生等五大类为本体模型的核心类。在此基础上，针对先秦典籍的特点，将 CIDOC CRM 中现有的类别层次及属性层次进行针对性的裁剪和扩充，进一步界定相关术语、概念、属性及适用对象范围在典籍这一具体情境中的确切含义。在军事、婚姻、外交、政治、民生等五大类的框架下，将典籍中蕴含的实物物件、符号物件、概念物件等对应的属性关系层次体系进行分面归纳，构建了面向典籍的本体模型[19-21]。

3.1.2 本体语义关系的抽取技术

学者们利用自然语言处理技术及机器学习技术，探索出了一系列本体语义关系抽取的方法[10]。然而，先秦典籍的语法和句法与现代汉语存在较大差异，导致目前在自然语言处理中的方法和技术不能直接应用于典籍本体语义关系的抽取中。以《左传》为例，先秦典籍具有句子长度较短、篇章语义主题不集中等问题。

语料资源也是本体语义关系抽取的重要基础，现代汉语积累了大量的标注语料，为本体构建奠定了良好的数据基础。但与现代汉语相比，囿于古汉语语法的难度以及数字化资源数量有限等因素，公开的典籍标注语料数量极少，这就给面向典籍的本体语义关系抽取带来了极大的难度。本文使用的数据来源是南京师范大学陈小荷研究团队标注的《左传》语料[21]，该语料对《左传》进行了人工分词和词性标注，是目前少有的先秦高质量语料。为了能够实现对《左传》中语义关系的抽取，本文仍然需要对其进行语义标注。由于先秦语法的特点，增加了标注的工作量和难度。

3.2 先秦典籍本文的标注方法

本体语义关系抽取从本质上而言,是抽取蕴含在文本中“主谓宾”三元组关系。基于此,本文将句法和角色结合,使用角色 BIO 标注方法对典籍文本进行标注。各标签含义见表 1,包括施事者(Agent)、受事者(Patient)、工具(Instrument)、处所(Location)和时间(Time)等。

表 1 BIO 标签含义

BIO 标注符号	含义
B-E1	施事者元素的第一个单词
I-E1	施事者元素除了第一个以外的其他单词
O-E1	施事者元素的唯一单词
B-E2	受事者元素的第一个单词
I-E2	受事者元素除了第一个以外的其他单词
O-E2	受事者元素的唯一单词
B-V	主题词元素的第一个单词
I-V	主题词元素除了第一个以外的其他单词
O-V	主题词元素的唯一单词
B-L	地点元素的第一个单词
I-L	地点元素除了第一个以外的其他单词
O-L	时间元素的唯一单词
B-T	时间元素的第一个单词
I-T	时间元素除了第一个以外的其他单词
O-T	地点元素的唯一单词
O	无意义词

按照角色 BIO 标注方法,依据先秦本体的语义框架对《左传》进行标注,以“仲慶父請伐齊師”为例,对应的标注数据如表 2 所示:

表 2 基于 BIO 标注方法的文本标注示例

词性	词组	标注
nr	仲慶父	O-E1
v	請	B-V
v	伐	I-V
ns	齊	B-E2
n	師	I-E2
w	。	O

根据标注的结果,谓词 V 对应本体模型中的属性类型,实施者 E1 对应本体对象属性(Object Property)的定义域(Domain),受试者 E2 对应本体对象属性的值域(Range)。因此,对于本体构建技术而言,获取大量角色 BIO 标注结果是实现本体实例抽取的重要基础。

4 面向先秦典籍的本体实例获取技术研究

针对先秦典籍在句法和句式上存在的特点,本文研究基于规则和条件随机场相结合的本体实例获取方法。从上文的分析不难发现,BIO 标注对于本体实例的获取至关重要,而 BIO 标注过程费时费力,本文探索利用条件随机场方法进行典籍语料标注。在 BIO 识别的基础上,利用谓词类型(下文称为“触发动词”)的语义关系识别本体实例的语义类型。

4.1 基于条件随机场的对象属性关系获取

对象属性关系主要是两个类别之间的属性关系,分别对应了典籍 BIO 标注中的实施类和受事类,而属性类型则由触发词的语义类型所决定。因此,为了获取大量的 BIO 标注结果,笔者将相关角色的获取看作一个序列标注问题,然后通过条件随机场模型获得相关的角色。

条件随机场模型(CRF)是 J. D. Lafferty 等[22]在 2001 年提出的条件概率分布模型,训练拟合目标为 P(Y|X),指是在输入随机变量 X 的条件下,根据特征模板构建特征函数,作为条件随机场的统计数据,进行训练拟合,预测输出随机变量 Y 的联合概率分布,并最终找到最高概率的最佳输出序列。在 CRF 算法中,选择合适的特征对训练结果有着巨大的影响。特征分为状态转移矩阵和观察序列特征。表 3 为训练数据中的一段特征和目标标签。

表 3 训练数据示例

特征	目标标签	当前词
秋	O-T	
,	O	
師	O-E1	<current token>
還	O-V	

在设定“O-E1”为当前词标签,“師”为当前词的情况下,表 4 根据特征模板中的 5 个特征做出特征说明。

表 4 特征模板说明

特征模板	特征	状态
U01:%x[－1,0]	,	当前词的前面第一个词
U02:%x[0,0]	師	当前词
U03:%x[1,0]	還	当前词的后面第一个词
U04:%x[－1,0]/%x[0,0]	,/師	前一个词到当前词的转移概率
B	O-E1/O	当前词标签和上一时刻标签

由于 CRF 算法已经较为成熟,本文不再赘述其实现过程,不同特征模板的识别效果将在实验测评部分具体展开。在先秦本体中,不同类别属性关系由触发词的语义类型决定。在获取了相应的角色后,通过句子所在的触发动词的语义类型来判别出角色所对应的本体属性类型。

4.2 触发动词的识别

通过对典籍文本的分析，不难发现句子中触发动词的语义往往决定了句子的语义类型。触发动词对于先秦知识本体中属性关系的识别至关重要，识别出触发动词的语义类型，再获取其对应的 BIO 角色，则完成了类别属性关系的获取。

为了构建本体属性关系对应的触发词语义类型，本文首先统计先秦文本中动词词频，根据先秦知识本体模型中的类别属性关系，通过手工的方式确定初始的触发动词集合，然后使用 BootStrapping 自扩展算法对触发词集进行扩充，进一步完善触发动词集合。

BootStrapping 是一种基于统计学知识的非监督学习机器学习算法[22]。它是在建立初始的核心词集的基础上，通过统计词与词之间的共现程度，计算相关度，设置一定的阈值，将高于该值的词与核心词集归为同一类，并加入到核心词集中，以此迭代训练，直到核心词集不再改变为止。具体算法是，首先以 tf-idf 值作为词的权重，选择权重在前 k 个的新词，或根据任务需求人工确认初始主题词集。其次，通过评估函数 T 计算每个候选词的分值，选择前列加入主题词集，生成新的主题词集，并且不断迭代。迭代条件可以是手工设置的迭代次数或者主题词集个数等。

$$T = \log_2 F(w,s) \times \frac{F(w,s)}{F(w)} \qquad \text{公式(1)}$$

公式(1)中的 s 和 w 分别表示核心词集和某候选词，F(w,s)表示某候选词与核心词集中的所有核心词共现频次的总和。

4.3 基于规则的本体数据属性关系获取

本体的数据属性关系(Data Property)不涉及两个类别之间的关系，如人物的基本特征，包括人名、性别、国别和官职情况等。根据对先秦文本进行分析，不难发现由于先秦典籍句式特点，使得这些基本属性关系在文本中呈现出一定的规律。因此，可以通过编写正则表达式的方式提取具有固有表述规则的属性。在手工标注的基础上，通过对样本进行分析，排除干扰因素归纳提炼规则，然后利用正则表达式抽取这些属性关系。具体的识别流程如下：

(1)对词性直接为 nr 的词语进行组成字和对应位置的分析，总结出 nr 的词与字的联系。

(2)对词性为 nr 的词语人名左右词的词性和词进行统计分析，找出包含基本属性关系的其他词汇。

(3)根据以上分析，获取人基本属性关系的构建规则。

(4)根据规则，构建正则表达式识别出所有符合要求的实例。

5 实验测评

5.1 测评方法

5.1.1 实验语料

笔者选取先秦时期的重要典籍《左传》作为实验语料[22]，该古汉语语料包含了《左传》分词、词性标注结果。在此基础上进行 BIO 角色标注，标注样例如表 5 所示：

表 5　标注格式样例

篇章	文本	词性	分词	标签
莊公十一年	冬，齊侯來逆共姬	n	冬	O-T
		w	，	O
		nr	齊侯	O-E1
		v	來	B-V
		v	逆	I-V
		nr	共	O-E2
		w	姬	O

经过手工标注及一致性检测，各标签的频次统计如表 6 所示：

表 6　标签－频次

	V	E1	E2	L	T
B	484	231	249	38	145
I	512	295	413	59	177
O	2 800	1 746	1 626	446	293

5.1.2 测评指标

本次实验采用准确率(Precision)、召回率(Recall)、F 值(Fscore)3 个评价指标作为实验结果的评价指标。

准确率(P)＝正确预测的标签个数/机器预测的全部标签个数＊100%

召回率(R)＝正确预测的标签个数/实际存在的所有标签个数＊100%

F 值(Fscore)＝2＊正确率＊召回率/(正确率＋召回率)＊100%

5.2 触发动词抽取实验

笔者对《左传》全文的动词进行统计分析，采用 Bootstraping 迭代方法进行触发词的获取。T2－婚娶(Marriage)、T3－生育(Bear)、T6－驻守(Garrison)、T7－功伐(Attack)、T8－会盟(Alliance)、T9－政治(Politics)、E67－诞生(Birth)、E69－死亡(Death)、T12－仕途(Career)等 6 个属性类型对应的触发动词集合抽取结果如表 7 所示：

表 7　语义类型-触发词对应表

事件类	主题词 1	词频 1	主题词 2	词频 2	阈值	迭代	扩展后触发词集	数量
T7-功伐(Attack)	伐	601	敗	228	0.56	7	[‘伐’,‘敗’,‘假道’,‘報’,‘帥’…]	491
E69-死亡(Death)	殺	444	死	362	0.5	5	[‘殺’,‘死’,‘抉’,‘朽’,‘縊’,‘轘’]	6
T8-会盟(Alliance)	盟	284	會	254	0.5	5	[‘盟’,‘會’,‘尋’,‘謀’]	4
T12-仕途(Career)	奔	255	爲	641	0.5	5	[‘奔’,‘爲’,‘御戎’]	3
T9-政治(Politics)	立	317	亡	230	0.5	5	[‘立’,‘亡’]	2
T2-婚娶(Marriage)	聘	123			0.5	5	[‘聘’,‘娶’]	2

5.3　基于规则的数据属性关系的识别测评

按照4.3节中的规则识别与匹配流程,对实验语料中人物相关语料进行规则提取实验,具体实验结果如下:

(1)对词性为“nr”的词本身进行分析。由表8可以看出,“子”和“公”都是左右的高频词。“子”在左侧取决于先秦人物的取名方式,在右侧主要是因为古代以“子”表示对人的尊称。“公”作为单字,经常代表君王,而当在左侧时常以“公子”出现,在右侧时常表示某公。除了这两个字外,左侧字常有国家的名称,右侧字常有家族辈分的名称。

表 8　人名首尾字频

左	频次	右	频次
子	1 195	子	1 519
公	493	公	1 097
晉	327	侯	713
叔	326	氏	615
季	273	伯	494
齊	256	叔	241
趙	252	王	198
鄭	231	孫	196
楚	206	父	170
…	…	…	…

(2)对词性为“nr”的词语左右边界词进行统计分析。如表9所示,w代表标点符号,v是动词,p是介词,都不属于人物名称提取的有效组成部分。而地名ns和普通名词n,可以和人名组成人物名称,是规则获取的重要信息线索词。

接下来,对人名左边界词性为ns和n的词进行分析,从表10可以看出词性为ns的词性一般都是国家的名称,而n一般是职位和家庭等词。

(3)根据上文的分析,找到人物名称的构建规则:[国家/身份]+[人名]。然后编写设计正则表示式进行提取,匹配结果见表11。

(4)由上一步的匹配结果可以看出,识别出来的结果的最后一个词,即为人名。如果检索结果只有一个词且词头是国家名,那么此国家名即为该人物所属国别。如果词尾是“公”“侯”“伯”,代表该人是君王,且性别为男;如果是“氏”,则判断性别为女。根据以上规则,统计结果见表12。

(5)将基于规则的抽取结果反馈到语料当中,进行半监督式的标注学习,补充从错误中学习的新规则。

表 9　人名左右词性频次

词性		词性说明	频次
左边词	w	标点	7 633
	v	动词	2 887
	ns	地名	585
	p	介词	568
	n	普通名词	524
	c	连词	381
	…	…	…
右边词	v	动词	6 657
	w	标点	2 895
	d	副词	738
	u	助词	714
	p	介词	578
	c	连词	474
	…	…	…

表 10　人名左侧词频

ns	频次	n	频次
晉	132	子	47
鄭	95	弟	27
楚	85	令尹	24
齊	68	公子	22
宋	65	大夫	20
衛	24	先君	19
衞	23	行人	14
陳	16	夫人	13
莒	11	大子	12
周	10	君	11
…	…	…	…

表 11　正则匹配示例

正则表达式	匹配结果
([^]*? /ns{0,1}){0,2}[^]*? /nr’	姜氏/nr
	司空/n 無駭/nr
	鄭/ns 公子忽/nr
	齊/ns 東宮/n 得臣/nr
	…

表 12 基于规则的匹配结果

text	E21－人物	T14－性别	T16－职位	T15－国别
高齡/nr	高齡			
鄭子元/nr	鄭子元			鄭
晉景公/nr	晉景公	男	君王	晉
樊氏/nr	樊氏	女		
楚/ns 令尹/n 子重/nr	子重		令尹	楚
命大夫/n 士蔑/nr	士蔑		命大夫	
晉/ns 冠氏/nr	冠氏	女		晉
…	…	…	…	…

实验结果如表 13 所示，从表 13 中可以看出，通过规则匹配后的 F 值性别最高，官职信息最低。通过对错误数据进行分析，主要是有些人物信息描述复杂，如果都要一一添加到规则里，会导致规则过于冗余且针对性太强。因此，在召回率和准确率之间存在较强的制约关系，如何泛化抽取规则值得进一步研究。

表 13 基于规则的实验结果评估

标签	准确率(%)	召回率(%)	F 值(%)
T14－性别	98.01	98.01	98.01
T15－国别	89.74	97.76	93.58
T16－职位	83.48	96.23	89.40

5.4 基于条件随机场的对象属性关系的识别测评

对经过触发词集筛选后的文本，使用条件随机场算法进行角色识别。本实验设计了 3 个模板进行试验对比，使用如下模板进行训练。

模板一：设置左右为 1 的窗口，只使用词特征进行训练。

模板二：在模板一的基础上加上词性特征。

模板三：在模板二的基础上加上词与词性之间的关系特征。

模板四：在模板三的基础上加上词的位置特征。

从表 14 的实验结果中可以看出，模板二加入词性特征后测试结果有了显著的提升。模板三中加入词与词性之间的关系特征，识别结果有一定的提升。加上位置信息后的模板四能有效提升识别效果。因此，选择特征模板四作为 BIO 角色识别的最终模板。

表 14 不同特征模板的识别效果

特征模板	准确率(%)	召回率(%)	F 值(%)
模板一	68.25	76.40	72.10
模板二	74.94	90.38	81.94
模板三	75.68	89.79	82.13
模板四	83.28	85.13	84.19

模板四的各类标签的测试结果如表 15 所示。其中 O-V、O-E1、O-E2、O-L、O-T、B-E1、I-E1 的 F 值均在 80% 以上。通过对错误数据进行分析，发现单字词的识别准确率高，而双字词及多字词识别的准确率较低。这是由于古汉语中单字词居多、训练样本分布不足，因此，可以尝试扩大训练数据集来提升实验结果。

表 15 基于模板四的本体实例抽取实验结果

实体标签	准确率(%)	召回率(%)	F 值(%)
B-L	75.00	40.00	52.17
I-L	83.33	43.48	57.14
I-V	71.35	67.56	69.40
B-V	71.26	68.51	69.86
I-E2	71.15	69.81	70.48
B-E2	71.43	74.26	72.82
O-L	73.91	77.78	75.80
O-E2	83.96	80.44	82.16
I-E1	86.75	78.26	82.29
O-V	85.07	84.74	84.90
B-E1	91.67	82.50	86.84
O-E1	86.92	87.30	87.11
O-T	96.52	94.87	95.69
B-T	96.36	96.36	96.36
I-T	97.22	98.59	97.90

6 结语

本文以《左传》文本为例，在构建了针对典籍文本内容的本体模型的基础上，探索针对先秦典籍文本的本体自动构建方法与技术。结合先秦典籍文本的特点，本文探讨了基于规则和条件随机场相结合的本体实例自动获取方法，通过实验表明，本文所提出的方法能够较好地从先秦典籍中抽取本体实例。该本体的构建能够较为全面地描述春秋社会的人物、事件相关信息，帮助研究者从线性的文本信息中挖掘隐性知识。

然而本文所提出的本体构建方法仍然还存在不足之处。一方面是触发动词的获取，未来可以通过外部词典辅助的方式提高触发词获取的准确率与召回率。另一方面，在本体实例抽取的过程中，只选取了词信息、词性信息和位置信息。在今后的工作中，可以通过提高句法分析的精度来获取更多的特征，进一步改善抽取效果。同时，也可以对如何利用深度学习技术提升本体实例的提取效果进行探索。

参考文献：

[1] 踪凡. 让古籍文献“活起来”[N]. 光明日报,2017-11-30(14).
[2] 夏翠娟,张磊. 关联数据在家谱数字人文服务中的应用[J]. 图书馆杂志,2016,35(10):26-34.
[3] 于彤,崔蒙,李海燕,等. ISO 技术规范“中医药学语言系统语义网络框架”的应用研究[J]. 中国医药导报,2016,13(4):89-92.
[4] 董慧,徐雷,王菲,等. 基于语义系统的中华史籍分析研究[J].

图书馆理论与实践,2015(4):1 – 5, 46.

[5] 陈小荷. 先秦文献的信息处理[M]. 北京：世界图书出版公司，2013.

[6] 欧阳剑. 面向数字人文研究的大规模古籍文本可视化分析与挖掘[J]. 中国图书馆学报,2016,42 (2): 66 – 80.

[7] 朱晓,金力. 条件随机场图模型在《明史》词性标注研究中的应用效果探索[J]. 复旦学报(自然科学版),2014,53(3):297 – 304.

[8] 刘浏,李斌,曲维光,等. 先秦词汇的时代特征自动获取及文献时代的自动判定[J]. 中文信息学报, 2013, 27(5):107 – 113.

[9] 于丽丽,丁德鑫,曲维光,等. 基于条件随机场的古汉语词义消歧研究[J]. 微电子学与计算机, 2009,26(10):45 – 48.

[10] 任飞亮,沈继坤,孙宾宾,等. 从文本中构建领域本体技术综述[J]. 计算机学报, 2019,42(3):654 – 676.

[11] WIMALASURIYA D C, DOU D. Ontology-based information extraction: an introduction and a survey of current approaches[J]. Journal of information science, 2010, 36(3):306 – 323.

[12] 王颖,张智雄,孙辉,等. 国史知识的语义揭示与组织方法研究[J]. 中国图书馆学报, 2015,41(4):55 – 64.

[13] THAKKER D, KARANASIOS S, BLANCHARD E, et al. Ontology for cultural variations in interpersonal communication: building on theoretical models and crowdsourced knowledge[J]. Journal of the Association for Information Science and Technology, 2017,68(6): 1411 – 1428.

[14] 周耀林,赵跃,孙晶琼. 非物质文化遗产信息资源组织与检索研究路径[J]. 2017,36(8):166 – 174.

[15] ISO technical committee 46 variations in interpersoation, subcomittee SC4 e 46 variations in interper. Information and documentation -- a reference ontology for the interchange of cultural heritage information[S]. ISO 21127:2014. Geneva:ISO, 2014.

[16] DOERR M. The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata [J]. AI magazine, 2003,24 (3):75 – 92.

[17] 顾栋高. 春秋大事表[M]. 北京：中华书局,1993.

[18] 童书业. 春秋左传研究[M]. 上海:上海人民出版社,2019.

[19] 陈小洁. 基于本体的《左传》战争知识地图构建研究[D]. 南京：南京农业大学,2018.

[20] 陈雅玲. 基于 CIDOC CRM 的先秦人物知识本体构建方法研究[D]. 南京:南京农业大学, 2019.

[21] CHEN X H, LI B, FENG M X, et al. Ancient Chinese corpus [M]. Philadelphia: Linguistic Data Consortium, 2017.

[22] LAFFERTY J D, McCALLUM A, PEREIRA F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data [C]// Proceedings of the 18th international conference on machine learning. San Francisco: Morgan Kaufmann Publishers Inc, 2001: 282 – 289.

[23] 吕云云, 李旸, 王素格. 基于 BootStrapping 的集成分类器的中文观点句识别方法[J]. 中文信息学报, 2013, 27(5):84 – 93.

作者贡献说明：

何琳:论文选题与框架设计,论文撰写及修改；

陈雅玲:论文撰写与算法实现；

孙珂迪:数据分析与数据处理。

Research on Ontology Building Methods of Chinese Ancient Books

He Lin　Chen Yaling　Sun Kedi

College of Information Science & Technology, Nanjing Agricultural University, Nanjing 210095

Abstract: [**Purpose/significance**] It is very helpful to build semantic ontology of Chinese ancient books for texting mining and text analysis of China history. However, there are lots of differences between ancient and modern Chinese in syntactic structure. The difference makes a lot of difficulties in Ontology Building of Chinese ancient books. [**Method/process**] This paper focused on ontology building methods of ancient Chinese books based on Natural language processing (NLP) technique. We designed the ontology model based on CIDOC CRM which is an international standard for the description of cultural heritages. Then we gave a solution to extract instances of the ontology automatically which is a hybrid method of regulation extraction and CRFs recognition based on the syntactic structure of Chinese ancient books. At last, we did an examination using one of Chinese ancient books called *Zuo Zhuan*. [**Result/conclusion**] The experiment results show that our method can improve the extraction precision of Ontology instances, which can enhance the efficiency of ontology construction from Chinese ancient books. This paper got 93% F-score on the testing of regular-based method, and 82.51% F-score on CRFs method using the best feature template. It also finds that it is important to use the characters of the position and part-of-speech of words to enhance the extraction of ontology instances in our methods.

Keywords: pre-Qin of Chinese ancient books　*Zuo Zhuan*　Ontology building　CRFs　Regulation matching method